Selecting Closest Vectors Through Randomization

Selecting Closest Vectors Through Randomization

Carl Bosley
and
Michael O. Rabin

TR-01-00

Computer Science Group
Harvard University
Cambridge, Massachusetts

Selecting Closest Vectors Through Randomization

Carl Bosley, M. O. Rabin

May 10, 2000

Abstract

We consider the problem of finding the closest vectors to a given vector in a large set of vectors, and propose a randomized solution. The method has applications in Automatic Target Recognition (ATR), Web Information Retrieval, and Data Mining.

1 Introduction

We consider the following problem. We are given a dictionary D of n vectors in no particular order, a special vector v_0, and a distance function f from the vector space to the nonnegative reals which takes the value zero iff $x = y$. We wish to find, for a given k, the k closest vectors in the dictionary to v_0, that is, the set of k vectors $v \in D$ which minimize $f(v_0, v)$.

Our approach is to take a random sample of m vectors from the dictionary, find the d-th closest vector v_d from those m, and then select all vectors v from the dictionary s.t. $f(v_0, v) \leq f(v_0, v_d)$.

The advantage of this method is that we only have to make one pass through the dictionary, examining $f(v, v_0) \forall v \in D$. We then have a much smaller set of vectors to work with, and we have made only one distance comparison for most vectors in the dictionary. Furthermore, by precomputing a single random sample for all queries, we can save even more time in doing repeated queries, since linear memory access is faster than sampling. Total search time should be $cn + O(k + \frac{n}{k})$, where cn is the amount of time to stream all vectors through once; this is $cn + O(\sqrt{n})$ when k is $O(\sqrt{n})$.

We now assume that all distances are unique; the sampling part of the problem then reduces to sampling m numbers randomly from $\{1, 2, \ldots, n\}$ and determining the d-th smallest.

Lemma 1. The probability that the d-th closest vector in a set of m randomly chosen vectors is further than the k-th smallest vector out of the n vectors is

$$S(n, m, d, k) = \sum_{i=0}^{d-1} \binom{m}{i} \left(\frac{k}{n}\right)^i \left(1 - \frac{k}{n}\right)^{m-i}.$$

Proof. The d-th closest vector of m is further than the k-th smallest element of n iff we have at most $d-1$ elements of m chosen from the smallest k elements of n. Each term of the sum represents the probability that of the m elements exactly i fall into this range, where $0 \leq i \leq d$.

Note that if m, k are small, given k, m, n, d we can use this sum to quickly compute the exact probability that if we set the cutoff to be the d-th smallest element of m and take all vectors nearer than that vector, we get at least k vectors.

Indeed we have

Pr(at least k vectors) $= S(n, m, d, k-1)$,

Pr(at most αk vectors) $= 1 - S(n, m, d, \alpha k)$,

Since we expect d to be very small, it is possible to use this formula to quickly calculate how large m should be to guarantee a very high chance of getting a large enough sample size and also a low chance of getting too large a sample size.

This should allow us to determine very quickly – in $O(d)$ time – whether a given m is large enough for a given n and k. Combined with binary search for m, this gives us an algorithm to compute the smallest value of m in $O(d \log m)$ time.

2 Approximating Probability of Failure

In this section we will derive an estimate for the probability our algorithm fails (returning fewer than k vectors or more than αk vectors). In practice we shall know n and k, so we shall not need to use this estimate, but in deriving it we see that probability of failure can be made close to 0 for large m.

Let $\beta = \frac{mk}{n}$. Then since $\lim_{x\to\infty}(1 - \frac{1}{x})^x = e^{-1}$, and the ratio of consecutive terms is $\frac{(m-i)k}{(i+1)(n-k)} \approx \frac{\beta}{i+1}$, we have

$$S(n, m, d, k) \approx e^{-\beta}\left(1 + \beta + \frac{\beta^2}{2} + ... + \frac{\beta^{d-1}}{(d-1)!}\right)$$

These sums are nice because we have removed the dependency on n, k (or αk), and m separately and replaced them by a single variable β.

We can get an even shorter approximation by bounding the value given by the series for e^β.

Let $R_d(x) = e^x - (1 + x + ... + \frac{x^{d-1}}{(d-1)!})$.

Then we have $R_d(x) = \frac{x^d}{d!}\left(1 + \frac{x}{d+1} + \frac{x^2}{(d+1)(d+2)} + ...\right) < \frac{x^d}{d!}\frac{1}{1-\frac{x}{d+1}}$ by geometric series. Similarly we can get a lower bound on the remainder this way.

This upper bound on the remainder allows us to get a close lower bound on the probability of selecting too few vectors – using this we find that

$$\Pr(\text{too few vectors}) < e^{-\beta}\left(\frac{\beta^d}{d!}\frac{1}{1 - \frac{\beta}{d+1}}\right).$$

3 Expected Value of Sample Size

We may wish to approximate how many vectors on average are returned by our method, so that we can try to fit them all in cache, for example. We shall calculate the expected sample size in this section.

We have that the expected value of the sample size is given by the sum

$$E(n, m, d) = \sum_{k=1}^{n} k\left(S(n, m, k, d) - S(n, m, k-1, d)\right) = \sum_{k=1}^{n} S(n, m, k, d).$$

For $d = 1$ this reduces to

$$\sum_{k=1}^{n}\left(1 - \frac{k}{n}\right)^m = n^{-m}\sum_{k=1}^{n} k^m.$$

In general we can calculate the difference between expected values of consecutive values of d:

$$E(n, m, d+1) - E(n, m, d) = n^{-m}\binom{m}{d}\sum_{k=1}^{n}(n-k)^d k^{m-d}.$$

It is known, and can be proven by induction on m, that the sum $1^m + 2^m + ... + (n-1)^m$ is equal to

$$\frac{1}{m+1}\sum_{k=0}^{m}\binom{m+1}{k}B_k n^{m+1-k},$$

where the B_i are the Bernoulli numbers, $B_0 = 1, B_1 = -\frac{1}{2}, B_2 = \frac{1}{6}, B_3 = 0, ...$, which are defined by the recurrence relation $\sum_{j=0}^{m}\binom{m+1}{j}B_j = 0, m > 0, B_0 = 1$.

It is also known that $\forall k > 0\ B_{2k+1} = 0$, and that $\forall k > 0$

$$\sum_{i=1}^{\infty} i^{-2k} = \zeta(2k) = (-1)^{n-1}\frac{2^{2n-1}\pi^{2n}B_{2n}}{(2n)!}.$$

Since $\zeta(2k)$ is close to 1, this gives a bound on B_{2k}.

Specifically we have, grouping together the 2^i-th term through the $(2^{i+1}-1)$-th term,

$$1 < \zeta(2k) < 1 + 2(2^{-2k}) + 4(4^{-2k}) + \cdots = 1 + 2^{-2k+1} + 4^{-2k+1} + \cdots = \frac{1}{1-2^{-2k+1}} = \frac{2^{2k-1}}{2^{2k-1}-1}.$$

Using this we get

$$\frac{(2k)!}{2^{2k-1}\pi^{2k}} < (-1)^{k-1}B_{2k} < \frac{(2k)!}{(2^{2k-1}-1)\pi^{2k}},$$

which implies

$$0 < |B_{2k}| < \frac{(m+1)^{2k}}{(2^{2k-1}-1)\pi^{2k}} < \left(\frac{m+1}{2\pi}\right)^{2k},$$

which decays quickly. From this we can find

$$E(n,m,1) = \frac{n}{m+1} + \frac{1}{2} + \frac{m}{12n} - \frac{1}{m+1}\sum_{k=4}^{m}\binom{m}{k+1}B_k > \frac{n}{m+1} + \frac{1}{2},$$

$$E(n,m,1) = \frac{n}{m+1} + \frac{1}{2} + \frac{m}{12n} - \frac{m(m-1)(m-2)}{720n^3} + \frac{1}{m+1}\sum_{k=6}^{m}\binom{m}{k+1}B_k < \frac{n}{m+1} + \frac{1}{2} + \frac{m}{12n},$$

which implies

$$\frac{n}{m+1} + \frac{1}{2} < E(n,m,1) < \frac{n}{m+1} + \frac{1}{2} + \frac{m}{12n}.$$

Now, we have (letting $D(n,m,d) = E(n,m,d+1) - E(n,m,d)$)

$$\begin{aligned}
D(n,m,d) &= n^{-m}\binom{m}{d}\sum_{k=1}^{n}(n-k)^d k^{m-d} && (1)\\
&= n^{-m}\binom{m}{d}\sum_{i=0}^{d}(-1)^{d-i}\binom{d}{i}n^i\sum_{k=0}^{m-i}\frac{1}{m+1-i}\binom{m+1-i}{k}B_k n^{m+1-i-k} && (2)\\
&= n^{-m}\binom{m}{d}\sum_{i=0}^{d}(-1)^{d-i}\binom{d}{i}n^i\sum_{k=0}^{m-i}\frac{1}{k}\binom{m-i}{k}B_k n^{m+1-i-k} && (3)\\
&= n^{-m}\left(\frac{n^{m+1}}{m+1} + \binom{m}{d}\sum_{k=0}^{m}\frac{1}{k}\binom{m-d}{k-1-d}B_k n^{m+1-k}\right) && (4)\\
&= n^{-m}\left(\frac{n^{m+1}}{m+1} + \sum_{k=0}^{m}\frac{1}{k}\binom{m}{k-1}\binom{k-1}{d}B_k n^{m+1-k}\right), && (5)
\end{aligned}$$

which is $\frac{n}{m+1} + O(\frac{1}{n^d})$.
So we have that

$$E(n,m,d) = \frac{nd}{m+1} + \frac{1}{2} + O(\frac{m}{n}) \approx \frac{nd}{m+1} + \frac{1}{2}.$$

4 Variance of Sample Size

We may further wish to solve several instances of the problem in parallel. In this case, in order to calculate the probability that the vectors fit in our cache, we may need to know not only the expected value, but also the variance of the sample size.

We have, letting $V(X)$ be the variance of a random variable X, and $E(X)$ the expected value, that $V(X) = E(X^2) - E(X)^2$.

Therefore the variance in the number of vectors chosen, which we shall label $V(n,m,d) = V(X)$, is approximately equal to $E(X^2) - (\frac{nd}{m+1} + \frac{1}{2})^2$.

Now, we have

$$E(X^2) = \sum_{k=1}^{n} k^2(S(k) - S(k-1)) = \sum_{k=1}^{n} (2k+1)S(k).$$

As before, we can write this as

$$E(n,m,d) + 2nE(n,m+1,d) = \frac{nd}{m+1} + \frac{1}{2} + \frac{2n^2d}{m+2} + n.$$

Hence we get that

$$\begin{aligned}
V(n,m,d) &= \frac{nd}{m+1} + \frac{1}{2} + \frac{2n^2d}{m+2} + n - E(n,m,d)^2 + O(\frac{m}{n}) && (6)\\
&= \frac{nd}{m+1} + \frac{1}{2} + \frac{2n^2d}{m+2} + n - \left(\frac{n^2d^2}{(m+1)^2} + \frac{nd}{m+1} + \frac{1}{4}\right) + O(m) && (7)\\
&= \frac{1}{4} + \frac{2n^2d}{m+2} + n - \frac{n^2d^2}{(m+1)^2} + O(m). && (8)\\
&\approx \frac{1}{4} + \frac{2n^2d}{m+2} + n - \frac{n^2d^2}{(m+1)^2}. && (9)
\end{aligned}$$